Dataset name: **2025-A001 - Pupil Census Ingest - Academic Years 2023/4 and 2024/5**

Stage 1: Preprocessing

No names are available for pupil census data

Total number of Input Records:	1,415,506
Total number of Input Persona:	770,231

Data requiring matching to the spine (i.e. new pupils)

New Input Records (not on existing read-through)	**187,328**
New Input Persons (not on existing read-through)	**125,944**

	N. Persons	**%**
Gender	125,944	
Valid Gender	125,944	
Day of Birth	125,944	100.0%
Valid Day of Birth	125,944	100.0%
Month of Birth	125,944	100.0%
Valid Month of Birth	125,944	
Year of Birth	125,944	100.0%
Valid YOB	125,875	99.9%
Postcode	125,944	100.0%
Valid Postcode	125,944	

Records with completed PII	125,944	100.0%
valid DOB, gender and postcode	125,875	99.9%
Duplicates	0	0.0%
Total unique input records	125,944	100.0%

Further pre-processing:

None

Dataset name: **2025-A001 - Pupil Census Ingest - Academic Years 2023/4 and 2024/5**

Stage 2: BigMatch Matching against the Indexing Spine

BigMatch is a linkage software program developed by the Statistical Research Division, U.S. Bureau of Census. More information:

Big Match: A Program for Extracting Probable Matches from a Large File for Record Linkage

The program is a linkage engine and implements traditional probabilistic record linkage methodology following the Fellegi-Sunter model for record linkage

BigMatch is designed to extract plausible matches from a large file using several blocking criteria without having to sort the file before each blocking run. Blocking is a commonly used technique in record linkage to minimise the number of comparisons between records. Records are grouped into blocks based on specified values that agree, for example instead of comparing all records, only records with the same sex are compared. Indexers select the most efficient blocks to perform the matching. This document contains results stratified by blocks. More information about blocking:

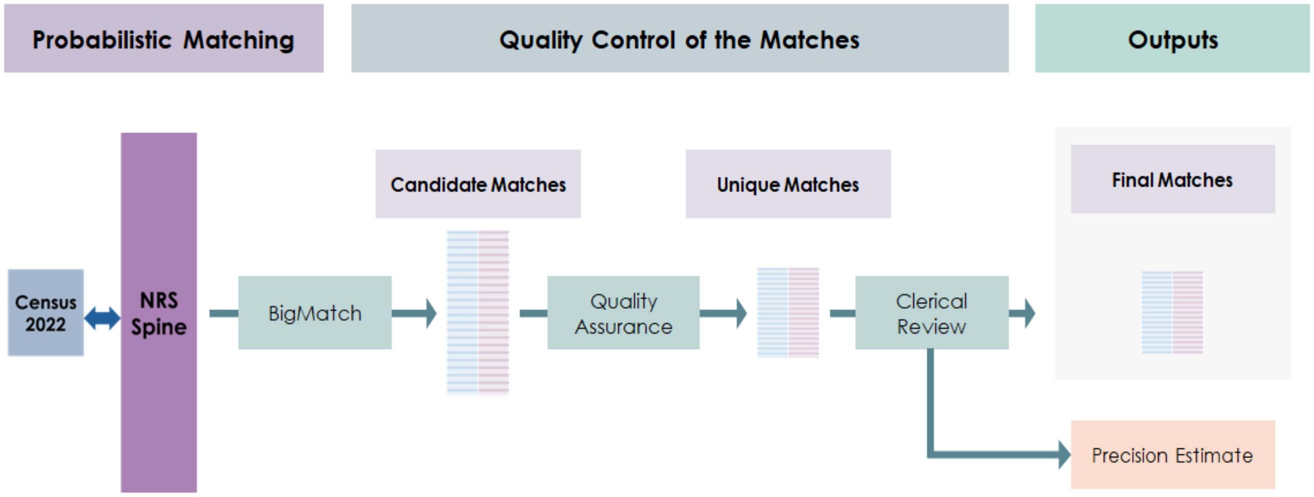
https://usc-isi-i2.github.io/papers/michelson06-aaai.pdf

BigMatch was run using the following hierarchical blocking criteria :

Block#	Block description	Final matches
1	Exact matches on DOB, sex and full postcode	119,405
2	Exact matches on DOB and full postcode	249
3	Exact matches on DOB, sex and first 6 characters of postcode	290
4	Exact match on year and month of birth, sex and full postcode	254
5	Exact match on year and day of birth, sex and full postcode	129
6	Exact match on month and day of birth, sex and full postcode	48
7	Exact matches on DOB, sex and first 5 characters of postcode	861
8	Exact matches on DOB, sex and first 4 characters of postcode	791
9	Exact matches on year of birth, sex and full postcode	31
Total		122,058

Dataset name: 2025-A001 - Pupil Census Ingest - Academic Years 2023/4 and 2024/5

Stage 3: Match Quality



Candidate Matches Total number of matches between external records and Spine IDs identified by BigMatch. This can include several records matching to the same Spine ID before any quality control.

Final Matches Competing matches are removed. The pupil data does not include names so clerical review will not help decide thresholds. Matches which are strongest for the pupil may have better matches for the matched spineID. Small numbers of links were the 'best' match cannot be decided have been broken.

Block#	N. Candidate Matches	Final Matches
1	120,469	119405
2	2,089	249
3	3,091	290
4	9,910	254
5	3,847	129
6	7,070	48
7	20,052	861
8	42,689	791
9	23,261	31
Total	232,478	122,058

Only matches which are unique (no equivalent competing matches) and exact are used for research projects.
All come from block 1 but not all matches from block 1 are unique.

Unique exact matches	116,898

Dataset name: **2025-A001 - Pupil Census Ingest - Academic Years 2023/4 and 2024/5**

Stage 4: Bias Analyses

The tables below indicate the number (and %) of matches in different categories of the data

Only new persons on the pupil census have been matched. **125,944**

Matches relate to unique exact spine matches

Data Completion

	Spine matching			
PII Completion	**UnMatched**	**Matched**	**Total**	**%Match**
Incomplete	68	1	69	1.4%
Complete	8,978	116,897	125,875	92.9%
Total	9,046	116,898	125,944	92.8%

Gender

	Spine matching			
Gender	**UnMatched**	**Matched**	**Total**	**%Match**
Male	4,637	60,061	64,698	92.8%
Female	4,409	56,837	61,246	92.8%
Total	9,046	116,898	125,944	92.8%

Year of birth

	Spine matching			
YOB	**UnMatched**	**Matched**	**Total**	**%Match**
2005 and earlier	83	47	130	36.2%
2006	123	367	490	74.9%
2007	292	890	1,182	75.3%
2008	313	1,230	1,543	79.7%
2009	313	1,318	1,631	80.8%
2010	299	1,506	1,805	83.4%
2011	306	1,513	1,819	83.2%
2012	267	1,854	2,121	87.4%
2013	268	1,908	2,176	87.7%
2014	274	2,044	2,318	88.2%
2015	276	2,221	2,497	88.9%
2016	303	2,375	2,678	88.7%
2017	514	5,100	5,614	90.8%
2018	2,519	47,146	49,665	94.9%
2019	2,689	44,250	46,939	94.3%
2020 or later	207	3,129	3,336	93.8%
Total	9,046	116,898	125,944	92.8%

These are only those pupils not seen before on the pupil census.

For older pupils, this is unexpected and might reflect people who have moved to Scotland or have changed their mode of education e.g. from private to state school.

Stage 4: Bias Analyses (continued)

Academic Year

	Spine matching			
Academic Year	**UnMatched**	**Matched**	**Total**	**%Match**
2023/4 and 2024/5	3,512	57,528	61,040	94.2%
2023/4 only	666	1,990	2,656	74.9%
2024/5 only	4,868	57,380	62,248	92.2%
Total	9,046	116,898	125,944	92.8%

SIMD

		Spine matching			
Decile	**Descriptor**	**UnMatched**	**Matched**	**Total**	**%Match**
1	Most deprived	1,287	14,813	16,100	92.0%
2		1,109	12,995	14,104	92.1%
3		891	11,691	12,582	92.9%
4		881	11,375	12,256	92.8%
5		758	10,371	11,129	93.2%
6		842	10,428	11,270	92.5%
7		833	11,293	12,126	93.1%
8		845	12,380	13,225	93.6%
9		680	11,213	11,893	94.3%
10	Least deprived	724	9,714	10,438	93.1%
Missing		196	625	821	76.1%
Total		9,046	116,898	125,944	92.8%

Dataset name: **2025-A001 - Pupil Census Ingest - Academic Years 2023/4 and 2024/5**

Stage 5: Spine matching summary

The category of matches we use for linkage projects is unique exact.
That is, there is only one matching spineID with exact match on sex, postcode and DOB

Probabilistic matching allows us to check there are no competing matches so we are confident the matches used are unique exact.

		% of cohort
Full pupil census ingest 2023/4 and 2024/5		
Persons	770,231	
Spine matches (unique exact)	723,611	93.9%
New pupils only from pupil census ingest 2023/4 and 2024/5	125,944	
Unique Exact spine matches	116,898	92.8%
For the previous update of the pupil census (2021-2023) new pupils gave a match rate of:		91.5%
Previously seen pupils from the latest ingest (2023/4 and 2024/5)	644,287	
Unique Exact spine matches for previously seen pupils	606,713	94.2%
Total number of Input Records (Full ingest 23/24 and 24/25):	1,415,506	
Remove duplicate storage key rows where demographic key missing		
Total number of output Records:	1,228,178	
Distinct storage keys	1,228,178	
Distinct demographic keys	1,228,178	
Combined persons from all pupil census (2007 to 2024/5)	1,769,930	
Spine matches (unique exact)	1,658,032	93.7%
Total number of output Records (from all pupil census):	12,446,100	
Distinct storage keys	12,446,100	
Distinct demographic keys	12,445,059	